KEEP CONFIDENTIAL

July 16, 2011

THE COMPUTATIONAL MAGIC OF THE VENTRAL STREAM: TOWARDS A THEORY

Tomaso Poggio⋆, † with appendices with and by Joel Leibo⋆ and Lorenzo Rosasco†
⋆ *CBCL, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA*
† Istituto Italiano di Tecnologia, Genova, Italy

ABSTRACT. I conjecture that the sample complexity of object recognition is mostly due to geometric image transformations and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. The most surprising implication of the theory emerging from these assumptions is that the computational goals and detailed properties of cells in the ventral stream follow from *symmetry properties* of the visual world through a process of unsupervised correlational learning.

From the assumption of a hierarchy of areas with receptive fields of increasing size the theory predicts that the size of the receptive fields determines which transformations are learned during development and then factored out during normal processing; that the transformation represented in each area determines the tuning of the neurons in the aerea, independently of the statistics of natural images; and that class-specific transformations are learned and represented at the top of the ventral stream hierarchy.

Some of the main predictions of this theory-in-fieri are:

- the type of transformation that are learned from visual experience depend on the size (measured in terms of wavelength) and thus on the area (layer in the models) – assuming that the aperture size increases with layers;
- the mix of transformations learned determine the properties of the receptive fields – oriented bars in V1+V2, radial and spiral patterns in V4 up to class specific tuning in AIT (eg face tuned cells);
- class-specific modules – such as faces, places and possibly body areas – should exist in IT to process images of object classes.

CONTENTS

1. INTRODUCTION

The ventral stream is widely believed to have a key role in the task of object recognition. A significant body of data is available about the anatomy and the physiology of neurons in the different visual areas. Feedforward hierarchical model, which are faithful to the physiology and the anatomy, summarize several of the physiological properties, are consistent with biophysics of corical neurons and achieve good performance in some object recognition tasks. However, despite the empirical and the modeling advances the ventral stream is still a puzzle: until now we do not have a broad theoretical understanding of the main aspects of its function and of how the function informs the architecture. The theory sketched here is an attempt to solve the puzzle. It can be viewed as an extension and a theoretical justification of the hierarchical models we have been working on. It has the potential to lead to more powerful models of the hierarchical type. It also gives fundamental reasons for the hierarchy and how properties of the visual world determine properties of cells at each level of the ventral stream. Simulations and experiments will soon say whether the theory has indeed some promise or whether it is nonsense.

As background to this paper, I assume that the content of past work of my group on models of the ventral stream is known from old papers [10–13] to more recent technical reports [2–6]. See also the section *Background* in Supp. Mat. [7].

1.1. **Recognition is difficult because of image transformations.**

The motivation of this paper is the conjecture that the "main" difficulty of (clutter-less) object categorization (say dogs vs horses) is due to all the transformations that the image of an object is usually subject to: translation, scale (distance), illumination, rotations in depth (pose). The conjecture implies that recognition – identification (say of a specific face relative to other faces) as well as categorization (say distinguishing between cats and dogs and generalizing from specific cats to other cats) – is easy, if the images of objects are rectified with respect to all transformations.

The conjecture is supported by new empirical evidence – so far just suggestive and at the anedoctal level – the "horse vs dogs" challenge (see Figure 4) – described in section 5.1. The figure shows that if we factor out all transformations in images of many different dogs and many different horses – obtaining "normalized" images with respect to viewpoint, illumination, position and scale – the problem of categorizing horses vs dogs is very easy: it can be done accurately with few training examples – ideally from a single training image of a dog and a single training image of a horse – by a simple classifier. In other words, the sample complexity of this problem is very low.

Additional support is provided by the following back-of-the-envelope estimates. Let us try to estimate whether the cardinality of the universe of possible images generated by an object originates more from intraclass variability – eg different types of dogs – or more from the range of possible viewpoints – including scale, position and rotation in 3D. Assuming a grain of a few minutes of arc in terms of resolution and a visual field of say 10 degrees, one would get $10^3 - 10^5$ different images of the same object from x, y translations, another factor of $10^3 - 10^5$ from rotations in depth, a factor of $10 - 10^2$ from rotations in the image plane and another factor of $10 - 10^2$ from scaling. This gives in the order of $10^8 - 10^{14}$ distingushable images for a single object. On the other hand, how many different distinguishable (for humans) types of dogs exist within the "dog" category? It is unlikely that they are more than, say, $10^2 - 10^3$. From this point

of view, it would be a much greater win to be able to factor out the geometric transformations than the intracategory differences.

Thus the key problem of recognition is to recognize – that is identify and categorize – from a single training image, invariant to geometric transformations. It has been known for a long time that this problem can be solved under the assumption that correspondence of enough points between stored model and new image can be computed. As one of the simplest such results, it turns out that under the assumption of correspondence, two training images are enough for orthographic projection (see [16]). Recent techniques for normalizing for affine transformations are now well developed (see Morel and Yu, 199 for a review). Various attempts of learning transformations have been reported over the years by Rao and Hinton among others [1, 8]. See for additional references the paper by Hinton [1].

The different goal here is to explore approaches to the problem that do not rely on correspondence and are plausible from the point of view of providing a plausible theory for the ventral stream.

My conjecture is then that *the main goal of the ventral stream is to factor out image transformations; furthermore invariance to transformations is be the main reason for the hierarchy and for the areas and cortical "patches" that are needed.*

1.2. **Plan of the paper.** In this introduction I just described the conjecture that the sample complexity of object recognition is mostly due to geometric image transformations, eg different viewpoints , and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. The most surprising implication of the theory emerging from these assumptions is that the computational goals and detailed properties of cells in the ventral stream follow from *symmetry properties* of the visual world through a process of correlational learning. The obvious analogy is physics: for instance, the main equation of classical mechanics can be derived from general invariance principles. In fact one may argue that a Foldiak-type rule determines by itself the hierarchical organization of the ventral stream, the transformations that are learned and the receptive fields in each visual area.

The first part of this paper deals with a few theoretical results that are independent of specific models. They are motivated by layered architectures "looking" at images, or at "neural images" in the layer below, through a number of small "apertures" corresponding to receptive fields, on a 2D lattice. I have in mind a *memory-based architecture* in which learning is just storing of image patches of patches of neural activation. This hypothesis could be made more complex but it is enough for our purposes.

There are four main sets of theoretical results:

(1) approximation of global transformations by local affine transformations
(2) recording transformed templates - *the templatebook* – provides a way to obtain an invariant signature for an object. This is the *invariance lemma*.
(3) the transformation "learned" at a layer depends on the aperture size. This is the *stratification theorem*.
(4) the spectral properties of the templatebook depends on the associated transformation(s).

From the assumption of a hierarchy of aereas with receptive fields of increasing size the theory predicts that the size of the receptive fields determines which transformations are learned during development and then factored out during normal processing; that the transformation

represented in an aerea determines the tuning of the neurons in the aerea; and that class-specific transformations are learned and represented at the top of the hierarchy.

The second part of this paper puts everything together in terms of a specific model – an extension of HMAX. It spells out the specific model and shows the implications of the theory of the first section. The connection is provided by two *linking theorems*:

(1) the max aggregation function maintains from layer to layer invariance of the signature of an image
(2) plausible forms of associative learning connect the spectral properties of the template-book to the tuning of simple cells at each layer in the model.

1.3. **Background.** In one of the early papers [10] we wrote:

It has often been said that the central issue in object recognition is the specificity-invariance trade-off: Recognition must be able to finely discriminate between different objects or object classes while at the same time be tolerant to object transformations such as scaling, translation, illumination, viewpoint changes, non-rigid transformations (such as a change of facial expression) and, for the case of cate- gorization, also to shape variations within a class.

and also

An interesting and non-trivial conjecture (supported by several experiments, of this population-based representation is that it should be capable of generalizing from a single view of a new object belonging to a class of objects sharing a common 3D structure such as a specific face to other views with a higher performance than for other object classes whose members have very different 3D structure, such as the paperclip objects. In a way very similar to identification, a categorization module say, for dogs vs. cats uses as inputs the activities of a number of cells tuned to various animals, with weights set during learning so that the unit responds differently to animals from different classes.

In the supermemo [11] I wrote:

Various lines of evidence suggest that visual experience – during and after development – together with genetic factors determine the connectivity and functional properties of units. In the theory we assume that learning plays a key role in determining the wiring and the synaptic weights for the S and the C layers. More specifically, we assume that the tuning properties of simple units – at various levels in the hierarchy – correspond to learning combinations of "features" that appear most frequently in images. This is roughly equivalent to learning a dictionary of patterns that appear with high probability. The wiring of complex units on the other hand would reflect learning from visual experience to associate frequent transformations in time – such as translation and scale – of specific complex features coded by simple units. Thus learning at the S and C level is effectively **learning correlations** *present in the visual world. The S layers' wiring depends on learning correlations of features in the image at the* **same time***; the C layers' wiring reflects learning correlations* **across time***. Thus the tuning of simple units arises from learning correlations in space (for S1 units the bar-like arrangements of LGN inputs, for S2 units more complex arrangements of bar-like subunits, etc). The connectivity of complex units arises from learning correlations over time, eg that simple units with the same orientation and neighboring locations should be wired together in a complex unit because often such a pattern changes smoothly in time (eg under translation).*

Since then we mainly focused on the hierarchical features represented by simple cells, on how to learn them from natural images and on their role in recognition performance. Here we focus on invariance and complex cells and how to learn their wiring, eg the domain of pooling.

As a consequence of this study, I have come to believe that I was wrong in thinking (implicitely) of invariance and selectivity as problems at the same level of importance. I now believe that the equivalence classes represented by complex cells are the key to recognition in the ventral stream. Learning them is equivalent to learning invariances and invariances are the crux of recognition in vision (and in other sensory modalities). I believe that the reason for multiple layers in the hierarchical architecture is the natural stratification of different types of invariances emerging from the unsupervised learning of the natural visual world with receptive fields of increasing size. In addition, the theory of this paper suggests that the tuning of the receptive fields in the hierarchy of visual areas is determined primarily by the transformations represented and discounted in each area.

2. THEORETICAL FRAMEWORK

In most of the paper I have in mind a hierarchical layered architecture as shown in Figure 3.1.1. However, this section is independent of the specifics of the hierarchical architecture. It deals with the computational problem of invariant recognition from one training image in a layered architecture. I also have in mind a computational architecture that is memory-based in the sense that it stores sensory inputs and does very little in terms of additional computations: it computes normalized dot products and max-like aggregation functions.

In this section I will discuss how geometric transformations of the image can be approximated by their "linear" term, eg affine transformations in $\mathbb{R}^2$. A memory-based approach to learning transformations is not restricted to affine transformations. I focus in this paper on the affine group to get some initial insight but it should be clear that this is just for convenience of analysis.

2.1. **Approximation of global transformations by a lattice of apertures.**

2.1.1. *Affine transformations in* $\mathbb{R}^2$. Let us assume – in this section – a x, y representation of images and transformations on them. In this representation, the components of the vector are the x, y coordinates of different features in an image. The features could be individual pixels *set in correspondence* across different images. A different representation that we will use in other parts of the paper is implicit and binary: we use 1 if a pixel is on and 0 otherwise. In this latter case, a vector corresponding to an image when displayed as a 2-D array is a binary image.

For each feature with coordinates x, y, we consider affine transformations defined as a 2×2 matrix

$$\Pi = \begin{pmatrix} a & b \\ d & e \end{pmatrix}. \tag{1}$$

Then an affine transformation with translations is

$$x' = \Pi x + t \tag{2}$$

with

$$x = \begin{pmatrix} x \\ y \end{pmatrix}$$

and

$$t = \begin{pmatrix} t_x \\ t_y \end{pmatrix}.$$

For a rotation of an angle θ the matrix $\mathbf{\Pi}$ is

$$\mathbf{\Pi} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}$$

It is possible to represent in a more compact way affine transformations (including translations) using homogeneus coordinates with the vector $\mathbf{v}$

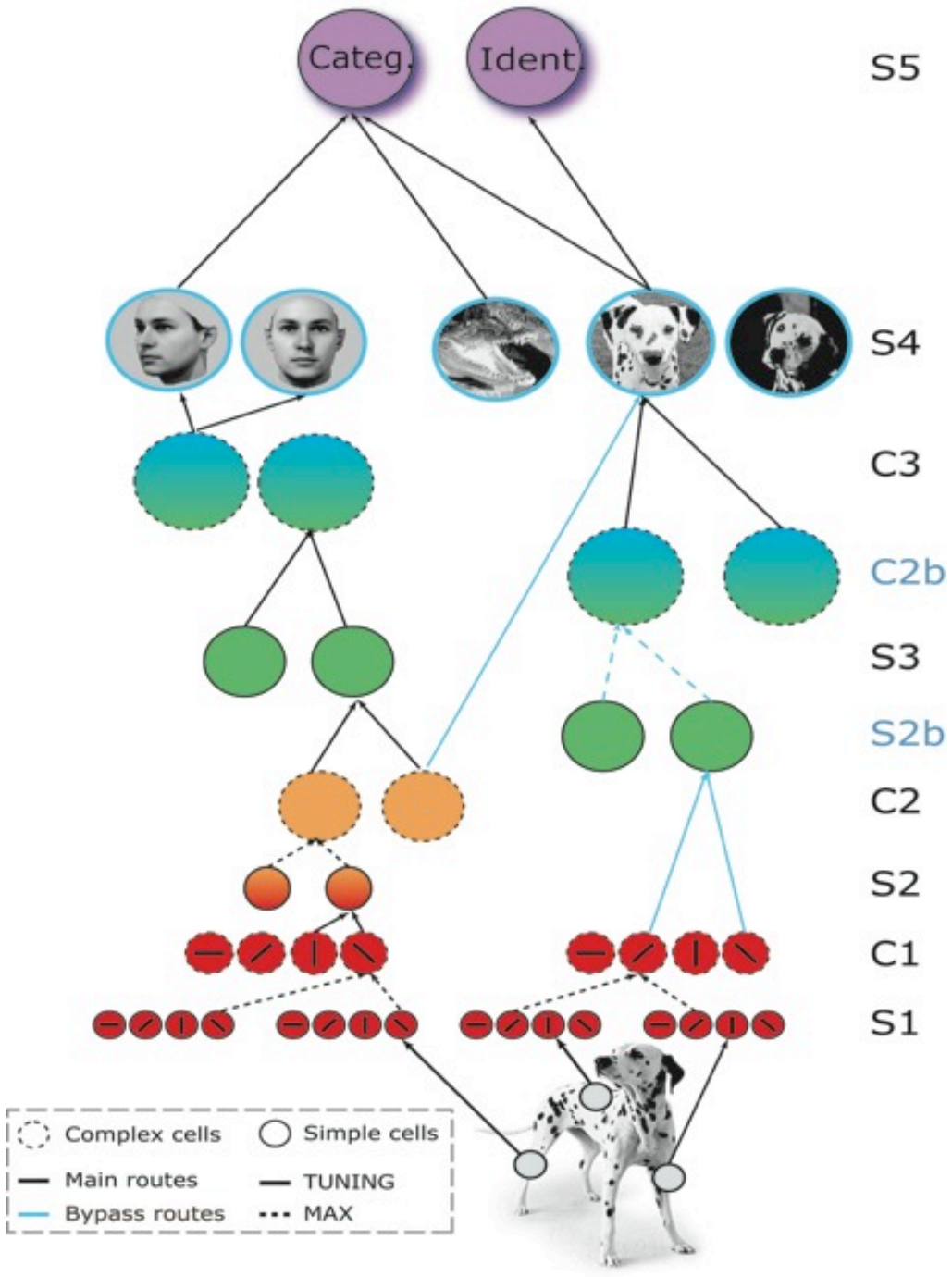

FIGURE 1. Hierarchical feedforward model of the ventral stream – a modern interpretation of the Hubel and Wiesel proposal (see [9]). The theoretical framework proposed in this paper provides foundations for this model and how the synaptic weights may be learned during development (and with adult plasticity). It also suggests extensions of the

$$x = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

and the 3×3 matrix $\mathbf{\Pi_{A'}}$ acting on it

$$\mathbf{\Pi}' = \begin{pmatrix} a & b & t_x \\ d & e & t_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

Thus $x' = \mathbf{\Pi}' x$.

Notice that the matrices $\mathbf{\Pi}'$ are representations of the affine group $Aff(2, \mathbb{R})$ which is an extension of $GL(2, \mathbb{R})$ by the group of translations in $\mathbb{R}^2$. It can be written as a semidirect product: $Aff(2, \mathbb{R}) = GL(2, \mathbb{R}) \times \mathbb{R}^2$ where $GL(2, \mathbb{R})$ acts on $\mathbb{R}^2$ in the natural manner (see Supp. Mat. [7]).

2.1.2. *Approximation of general transformations in* $\mathbb{R}^2$. Assume that there are several features in the image (in the limit these features may be pixels in correspondence). Then the image can be represented as a vector

$$\begin{pmatrix} x_1 \\ y_1 \\ 1 x_2 \\ y_2 \\ 1 \cdots \\ x_N \\ y_N \\ 1 \end{pmatrix} \tag{4}$$

Assume the same affine transformation is applied to the whole vector. Then Π is

$$\mathbf{\Pi} = \begin{pmatrix} A & 0 & \cdots & 0 \\ 0 & B & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & Z \end{pmatrix} \tag{5}$$

where $A, B, \cdots, Z$ have the form of equation 3. If the same affine transformation is applied everywhere then the 2×2 blocks is such that $A = B = Z$ (this is the case we call *globally affine*).

2.1.3. *The local affine approximation lemma.* Consider the array of Equation 5. Assume In the limit, if there are different affine transforms for each patch of the image then it is possible to approximate any global transformation of the image arbitrarily well for patch size going to zero (and increasing number of patches) if apprpriate (different) affine transformations are used for each patch. I conjecture the following approximation property:

Lemma 1. *Local affine transformations $Aff(2, \mathbb{R})$ on image patches can approximate arbitrarily well any smooth transformation in $\mathbb{R}^2$ of the image by increasing the number of patches and decreasing their size.*

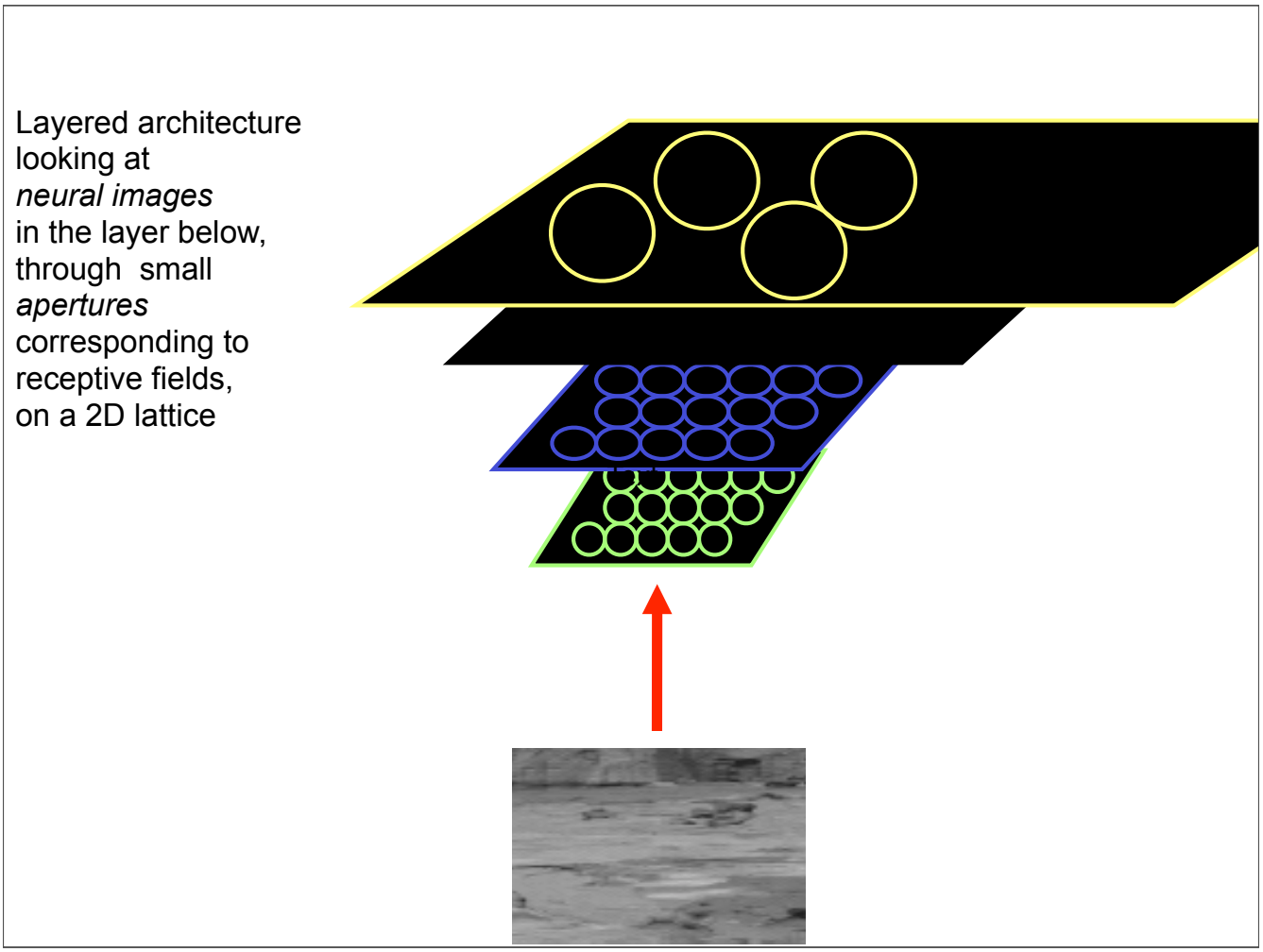

FIGURE 2. A layered architecture with apertures of increasing size (in reality overlapping).

Proof sketch If $\mathbf{x}_0$ is a point at the center of the patch and the transformation $Tf(\mathbf{x}) = f(T\mathbf{x})$ and T is differentiable at $\mathbf{x}_0$ then its derivative is given by $J_T(\mathbf{x}_0)$. In this case, the linear map described by $J_T(\mathbf{x}_0)$ is the best linear approximation of T near the point $\mathbf{x}_0$, in the sense that

$$T(\mathbf{x}) = T(\mathbf{x}_0) + J_T(\mathbf{x}) \cdot (\mathbf{x} - \mathbf{x}_0) + o(||\mathbf{x} - \mathbf{x}_0||). \tag{6}$$

The affine approximation lemma suggests the following conjecture

Conjecture 1. *A set of local "receptive fields" on a lattice are optimal for representing and correcting small global non uniform deformations of the image.*

The conjecture does not make sense without additional conditions. It may then be a justification for the evolution of architectures consisting of many local apertures, as a first layer.

Later I will argue that higher layers may be needed for larger but more constrained transformations.

2.2. **Templatebooks and Invariant Signatures.**

2.2.1. *Signatures of images and Johnson-Lindenstrauss.* As human we are estimated to be able to recognize in the order of 50K object classes through single images, each one with a dimensionality of 1M pixels (or ganglion cells axons in the optical nerve).

Since the goal of visual recognition in the brain is not reconstruction but identification or categorization, a representation possibly used by the ventral stream and suggested by models such as Figure 3.1.1 is in terms of an overcomplete set of measurements on the image, a vector that we call here *signature*. The components of this vector are the *normalized dot products* of the image (or image patch) f w.r.t. a set of *templates* $\tau_{\mathbf{i}}$, $i = 1, \cdot, D$, which are image patches themselves. We call the set of templates a *templateset*. More formally:

Definition 1. *The signature of* $\mathbf{f}$ *wrt the templateset* $\tau_{\mathbf{i}}$ *is the set* $K(f, \tau_i)$ *where* K *is the normalized kernel*

$$K(g,h) = \frac{g \circ h}{(|g \circ g||h \circ h|)^{1/2}}.$$

I am thinking of a recursive computation of signatures, from layer to layer from small patches to large ones. In general, signatures at various levels may be used by a classifier.

An empirical example and historical motivation for this approach is OCR done via intersection of letters with a random, fixed set of lines and counting number of intersections. A more mathematical motivation is provided by a theorem due to Johnson and Lindenstrauss. Their classic result asserts that *any set of n points in d-dimensional Euclidean space can be embedded into k-dimensional Euclidean space where k is logarithmic in n and independent of d so that all pairwise distances are maintained within an arbitrarily small factor.*

2.2.2. *Templatebooks.* Consider the templateset τ as a column vector consisting of D templates images, which can be thought as D templates (eg different patches of an image, corresponding for instance to different parts of the same object), observed at the same time $t = 1$:

$$\tau = \begin{pmatrix} \tau_{1,1} \\ \tau_{2,1} \\ \cdots \\ \tau_{D,1} \end{pmatrix}.$$

Assume now that the image of the object undergoes a geometric transformation due to motion (of the camera or of the object). For simplicity we will assume that the transformation can be well approximated locally (eg within the receptive field of the complex cells) by an affine transformation in R^2. We will denote as $\tau_{1,1}$, $\tau_{1,2}$, ...,$\tau_{1,N}$ the images of the same patch at frames (eg instants in time) $t = 1, \cdots, N$. Thus each row of the matrix $\mathbb{T}$ below represents the whole templateset τ at different times of a transformation:

$$\mathbb{T} = \begin{pmatrix} \tau_{1,1} & \tau_{1,2} & \cdots & \tau_{1,N} \\ \tau_{2,1} & \tau_{2,2} & \cdots & \tau_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ \tau_{D,1} & \tau_{D,2} & \cdots & \tau_{D,N} \end{pmatrix}.$$

We call the matrix $\mathbb{T}$ a *templatebook* corresponding to one of the templatesets and representing a certain transformation. Each row of the template book $\mathbb{T}$ corresponds to a set of simple cells, see Figure 3, pooled by the same complex cell. The set of simple cells are similar to frames in a video associated with a complex cell. Each row of $\mathbb{T}$ is a list of the simple cells pooled by a complex cell; each row corresponds to a different complex cell. Each row is the transformation of one template. There are different templatebooks, each for a different transformation.

2.2.3. *The invariance lemma.* Consider the signature vector $\mathbf{s}$ corresponding to an image patch f with respect to a templateset τ, as defined earlier. Thus $\mathbf{s} = (f \cdot \tau_1, f \cdot \tau_2, \cdots, f \cdot \tau_n)$, where

$$f \cdot t_i = \frac{\int f(x,y)\tau_i(x,y)dxdy}{[(\int f(x,y)dxdy)(\int \tau_i(x,y)dxdy)]^{1/2}} \tag{7}$$

Consider now geometric transformations $Tf(x,y) = f(u,v)$. We call the transformation *uniform* if the Jacobian $J(x,y) = constant$. As a major example T may correspond to affine transformations on the plane eg $\mathbf{x}' = A\mathbf{x} + \mathbf{t_x}$ with A a nonsingular matrix.

Then the following lemma holds

Lemma 2. *The signature of f with respect to set of templates $\tau_1, \cdots, \tau_n$ is equal to the signature of Tf w.r.t. the set of templates $T\tau_1, \cdots, T\tau_n$ for uniform geometric transformations.*

Proof sketch It is enough to consider the effect on one of the coordinates.

$$f \cdot \tau_i = \frac{\int f(x,y)\tau_i(x,y)dxdy}{[(\int f(x,y)dxdy)(\int \tau_i(x,y)dxdy)]^{1/2}} = \tag{8}$$

$$= \frac{\int f(u,v)\tau_i(u,v)|J(u,v)|dudv}{[(\int f(u,v)|J(u,v)|dudv)(\int \tau_i(u,v)|J(u,v)|dudv)]^{1/2}} =$$

$$= \frac{\int f(u,v)\tau_i(u,v)dudv}{[(\int f(u,v)dudv)(\int \tau_i(u,v)dudv)]^{1/2}} = Tf \cdot T\tau_i$$

The invariance lemma implies that independently of the templates – and how selective they are – the signature they provide can be completely invariant to a geometric transformation which is uniform over the pooling region. We will see later an architecture for which signatures are invariant recursively through layers. The actual templates themselves do not enter the argument: the set of similarities of the input image to the templates need not be high in order to be invariant.

2.2.4. *Invariant recognition from one training example.* Let us call a templatebook *complete* relative to transformation T, if for any uniform transformation T and for any entry $\tau_{i,j}$ there is a k s.t. $\tau_{i,j} = \tau_{i,k}$. The following obvious statement implies that f can be recognized from a single example, independently of the unknown but uniform transformation.

Theorem 1. *Assume there is a single training image f of an object. Assume that a new image Tf is given (with unknow T). If a complete templatebook is available in memory – from previous observations of even a single unrelated object – obtained under the same uniform transformation T, then the signature of f is independent of T.*

Proof sketch The proof follows from the Lemma and from appropriate conditions on the length of the templateset. The formulation of the theorem should be given in terms of bounds on probability of error.

2.3. **Stratification of and peeling off transformations.** Let me describe the set of ideas that motivate the analysis of this section. I have given some results and observations on how images could be recognized from a single training image independently of transformations by learning implicitely a transformation in terms of a templatebook. I assume now that this is done within the hierarchical architecture of the ventral stream. From this point of view it seems natural to learn and discount transformations (the one-step-process of Part I) – in a sequence, from simple and local transformations to complex and less local ones. The sequential aspect of learning transformations in a sequence of layers correspond to the term *stratification*, while the sequential use of the layers, one after the other at run time, corresponds to the term *peeling off*. In this section I conjecture that stratification appears because the kind of transformations that are learned depend on the aperture (eg the receptive filed size of a complex cell), which increases from V1 to IT. In other words I conjecture that receptive field size determines complexity (eg type) of transformations[1].

2.3.1. *Apertures and learnable invariances.* The question to be explored is whether the size of the aperture, eg of the receptive field, determines which affine transformations can be learned from general image sequences. The intuition is that for a small aperture, measured in terms of spatial frequencies of the input "images" (which can be neural "images" provided by the previous layer), translations should emerge first, followed by scalings and rotations, followed by full affine transformations, followed by class-specific transformations. The qualitative reasons for this argument are

- one corresponding point over two frames is enough for estimating the two translation parameters
- one additional point is needed to estimate the rotation angle or the scaling
- three points are needed to estimate the 6 affine parameters
- more points require larger aperture for fixed amount of noise in the measurements
- estimation of rotation and scale requires first estimation of the center of rotation and the focus of expansion, respectively. It is natural, therefore, to estimate translation first.

Notice that

- the aperture size limits the size of translations that can be "seen" and estimated from a single aperture (eg a single receptive field or cell)
- there is a tradeoff between large aperture for more robust estimation of larger transformations and the correspondence problem which increases in difficulty (I expect this is equivalent to using correlation between the two frames to estimate parameters)
- growing aperture size corresponds to increasingly complex features (from 1-point features to 3-point features)

A radially symmetric aperture (a disk) correspond to convolution in the Fourier domain with $J_1(\omega r)$, where J_1 is the Bessel function of the first kind. A large aperture corresponds to an increasingly delta-like Fourier transform; a small aperture corresponds to broader and broader

[1]Maha contributed to the arguments of this section in a few discussions.

enevlope and more and more "blurring" (in the frequency domain). The same reasoning can be repeated with a radially symmetric spatial Gaussian modeling the "aperture". It is also important to remember that we consider affine transformations which are *uniform* within an aperture (eg within the receptive field of one cell). Clearly, a large range of complex, non-uniform *global* transformations of the image can be approximated by local affine transformations or even just by local translations.

2.3.2. *Stratification conjecture.* Consider, given two or more frames observed through a small aperture of size r, the task of estimating an affine transformation. I conjecture that the following theorem (or some closely related statement) holds:

Theorem 2. *Assume a fixed amount of measurement noise. For aperture size increasing (from zero) the first transformation that can be estimated reliably above a fixed threshold, is translation. For increasing aperture size, the last transformation whose parameters can be estimated reliably above a fixed threshold, is full affine. The condition number of the problem is* 1 *for isotropic scaling; it can grow arbitrarily large depending on the asymmetry of the scaling.*

Proof sketch

An informal argument is the observation that estimation of translation requires estimating 2 numbers with condition number equal to 1. Rotation requires estimating 3 numbers (pure rotation plus translation) with condition number equal to 1. Scaling requires estimating 4 numbers (asymmetric scaling plus translation) – and the condition number could be bad. Therefore translation requires the least amount of bits, followed by rotation, followed by (asymmetric scaling), followed by full affine. Assume also that increasing aperture size corresponds to increasing number of bits.

I conjecture that the first layer of the model learn (small) translations and later layers learn mixtures of rotations and scaling and (larger) translations. Especially in the final layers the structure and the spectral properties of the templatebooks may depend on the natural statistics of image transformations. Simulations are needed to find out how many layers are needed to learn the full affine transformations over a range of translations. Notice that small apertures can only learn and represent small translations; larger ones are taken into account by larger apertures.

2.4. **Spectral properties of the templatebook.** In this section I will show that the spectral properties of the templatebook are determined by the transformation implicitely coded in it. The reason for characterizing the spectrum is that a class of natural learning rules for simple cells implies that their receptive field is determined by the spectral properties of of $\mathbb{T}$. This will be shown by the second linking theorem in section 3.1. Thus *transformations learned at a certain layer - for instance depending on aperture size - imply spectral properties of the templatebook which in turn imply tuning of receptive fields of associated simple cells.*

The analysis here assume a nongenerative scheme: transformation are learned implicitely by storing sequences of transformations of templates (as models described later do). There is however a possible "generative" version of the scheme outlined here. Each layer may transform the input "image" – instead of simply encoding invariant signatures. In this case the transformation operator – a kernel matrix Π – could be learned by an associative memory module which associates a templateset $x = \tau$ with its transformed version $y = T\tau$, both observed as frames of

a "video" recorded during a transformation. Thus Π represents the transformation between x and y and can transform x into y. I describe this "generative" scheme in the Supp. Mat. [7]).

I first recall some observations on affine tranformations.

Decomposition of affine transformations

There is another related decomposition of affine transformations, called the RQ decomposition. An homogeneous matrix A' can be decomposed as

$$A' = MK$$

where M is an orthogonal rotation matrix and $K = LS$ is an upper triangular matrix, L is a translation matrix and S is a shear and scale matrix. Thus

$$M = \begin{pmatrix} cos\theta & sin\theta & 0 \\ -sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$L = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix}$$

$$S = \begin{pmatrix} s_x & k & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$K = \begin{pmatrix} s_x & k & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{pmatrix}$$

.

We now turn to an analysis of the spectral properties of affine transformations.

Spectral properties of the translation operator in $\mathbb{R}^2$

The eigenfunctions depend on the representation we use for images. The standard representation is in terms of x, y coordinates of corresponding features or points in images. In this explicit representation of images as vectors of x, y coordinates, translations cannot be mapped to matrices acting on the vectors, unless I use homogeneous coordinates (see section 2.1.1). As we will see, this representation translations do not commute with scaling and rotation; scaling and rotation commute which other only if the scaling is uniform.

It is well known that the eigenfunctions associated with the translation operator (in R^2 in our case) are the complex exponentials. The informal argument runs as follows. Consider the translation operator acting on functions $\phi(x)$ in L_2 defined by $T_{x_0}\phi(x) = \phi(x - x_0)$. The operator $T_{x_0} = e^{-x_0 \frac{d}{dx}}$ is unitary and forms a representation of the additive group. The definition leads to a functional eigenvalue equation

$$\phi(x - x_0) = \lambda\phi(x)$$

with solutions (see Supp. Mat. [7]) $\phi(x) = e^{i\omega x}$.

Spectral properties of the scaling and rotation operators

The eigenfunctions of rotations and scaling are complex exponentials in polar coordinates. In other words $\phi(x, y) = \rho e^{i\theta}$ is a solution of the eigenvalue equation for the rotation operator R

$$R_{\psi_0}\phi = \lambda\phi$$

with $\lambda = e^{i\psi_0}$, and similarly for the scaling operator, where the eigenvalue is real

Compositions of transformations

Assume the semidirect product $Aff(2, \mathbb{R}) = GL(2, \mathbb{R}) \times \mathbb{R}^2$ for a composite transformation that I have introduced earlier. Let us focus on the linear transformations represented by a two-by-two matrix A, neglecting translations. A can be decomposed using SVD as

$$A = U\Sigma V^T$$

where all matrices are $2x2$, Σ is diagonal and U and V are orthogonal. Thus any affine transformation represented in this way is decomposed into a rotation followed by asymmetric scaling followed by a rotation. It follows that the condition number of A is 1 if scaling is isotropic and larger than 1 otherwise. It is possible to consider a sequence of transformations such as for instance scaling and rotation and analyze it in terms of the SVD decomposition.

Gabor wavelets

The windowed Fourier transform (WFT) and the inverse are

$$F(\omega, a) = \int dx f(x)G(x - a)e^{-i\omega x} \tag{9}$$

$$f(x) = \frac{1}{||G^2||}\int d\omega da G(x - a)F(\omega, a)e^{i\omega x} \tag{10}$$

An examination of the first equation shows that $F\omega, a)$ is the Fourier transform of $f(x)G(x - a)$, that is *the pattern $f(x)$ "looked at" through a Gaussian window $G(x)$ centered at a.* Since Fourier components emerge from translation, this implies that Gabor wavelets of the form $G(x)e^{-i\omega x}$ emerge from translations of $f(x)$ modulated by $G(x)$.

The previous argument is for a single (Gaussian, though this is not required) aperture centered in a. One can ask whether $f(x)$ can be represented in terms of several windows spanning a lattice in a space. The answer is affirmative.

In addition, it is possible to show that Gabor wavelets (with actually σ depending on ω) are a representation of the similarity group (in our case in R^2). Stevens [15] develops an interesting and detailed argument for Gabor receptive fields in V1 to be implied by "invariance" to translations, scale and rotations. His Gabor wavelets have the form

$$\phi_{\theta,\xi,\eta}(x, y) = e^{-\frac{(x_\theta - \xi_\theta)^2}{2\sigma_x^2}} e^{-\frac{(y_\theta - \eta_\theta)^2}{2\sigma_y^2}} e^{i2\pi x_\theta \omega}$$

where the center of the receptive field is ξ_θ, η_θ, the preferred orientation is θ and σ_x^2 and σ_y^2 are proportional to $frac1\omega^2$.

This family of 2D wavelets, and their 2D Fourier transforms, is each closed under the transformation groups of dilations, translations, rotations, and convolutions.

Gabor wavelets as diagonalizing templatebooks acquired under translation through a Gaussian window

The argument is about the spectral content of a templatebook acquired by recording videos of translation through a Gaussian window. Each row of the templatebook corresponds to a matrix in which each column is the image of the same patch translated. The matrix has therefore the structure of a Toepliz matrix whose spectral properties are closely related to circulant matrices. I describe the analysis for circulant matrices. This is what I would obtain from each row of the template book if the patterns moving behind the Gaussian window would be periodic (eg have the geometry of a thorus). In this case, the DFT diagonalizes the circulant matrix X. Thus

$$F^T X F = \Lambda,$$

where Λ is a diagonal matrix. In particular, a column of the matrix, which is an image "looked at" through a Gaussian window, can be represented as $GI = G \sum c_l e^{i\omega l x} = \sum c_l G e^{i\omega l x}$ thus in terms of Gabor wavelets.

The Supp. Mat. [7] part describes additional known results that extend the math of this section – for instance about the conditions on the lattice of the "apertures" to ensure good global representations.

Lie algebra and Lie group

The Lie algebra associated with the Lie group of affine transformations in $\mathbb{R}^2$ has as an underlying vector space the tangent space at the identity element. For matrices A the exponential map takes the Lie algebra of the general linear group G into G.

Thus a transformation T of x parametrized by s can be represented as $T = e^{As}$. Notice that if A is symmetric then $A = U \Lambda U^T$ and

$$T = e^{As} = e^{U \Lambda U^T s} = U e^{\Lambda s} U^T. \tag{11}$$

and thus *the spectrum of A and the spectrum of T coincide*. This is not true if A is not symmetric.

Summary As we saw, it is "easy" to characterize analytically what the spectral properties are for pure translations, rotations and scalings. It is not however as simple to characterize the spectral properties for mixture of transformations – that occurs for larger apertures and large enough transformations. Simulations (ongoing) will be necessary especially for characterizing the *spectral properties of transformations encoded in templatebooks acquired in a hierarchy of lattices of Gaussian apertures.*

3. A MODEL: DEVELOPMENT, LEARNING, COMPUTATION

3.1. **Linking Theorems.**

3.1.1. *Invariant aggregation functions.* I have in mind the model of Figure and the work around it. Let us assume that f is the image or the neural response of the layer below in the hierarchy. For each complex cell there is a set of templates – that is a template and its transformations over which the complex cell is pooling. Let us use each row in the matrix below as the vector of measurements relative to the transformations of one object-patch, listed in the row in the order in which they appear during the transformations. So the column index effectively runs through time and through the simple cells pooled by the same complex cell:

$$S = \begin{pmatrix} f \circ \tau_1^1 & f \circ \tau_1^2 & \cdots & f \circ \tau_1^n \\ f \circ \tau_2^1 & f \circ \tau_2^2 & \cdots & f \circ \tau_2^n \\ \cdots & \cdots & \cdots & \cdots \\ f \circ \tau_m^1 & f \circ \tau_m^2 & \cdots & f \circ \tau_m^n \end{pmatrix}.$$

The circuit needs to pool from each set of simple cells into a number, so that it provides to higher layer a signature – a vector of m components (as many components as complex cell).

To accomplish the model (for instance HMAX) uses an *aggregation function* such as a max or an average or a σ-oidal function operating on each element of the matrix – for σ – or along each row, that is over the column index – for the max.

We need to make sure that the aggregation function is indeed invariant. We can prove this using the invariance lemma 2 in the following way.

Assume the following *max learning rule*: *Assume that templatebooks have been acquired during development. The rule for acquiring the signature of a new image f at the top level is to use as aggregation function the max and to select at each level for each complex cell the value $f \circ t^* = max_{t \in T} f \circ t$ (this is the way HMAX works now).*

This means that at runtime the max will always (in the noiseless sitaution) choose the correct simple cell (corresponding to t*) and provide the same value – independently of T. Across complex cells this says that with high probability the signature is invariant to T from layer to layer (for uniform T). We have the following *aggregation theorem*:

Theorem 3. *If at some level the set of templates is complete then the max of the signature of f is invariant to affine transformations of f, that is $signature(f) = signature(Tf)$.*

Sketch of the proof By assumption the aggregation function chooses for an element of the signature at the higher level the max over templates t (see rule above), providing as component of the higher level signature $f \circ t^*$. After this choice is made during learning, assume that the new image to recognize at run time is $f' = Tf$. We claim that $max_{t \in T} f' \circ t = f \circ t^*$. Assume the opposite eg $max_{t \in T} f' \circ t \geq f \circ t^*$. This implies that $max_{t \in T} f \circ t \geq f \circ t^*$, which contradicts the assumption. Q.E.D.

3.1.2. *Learning rule and receptive fields.* The algorithm outlined earlier in which transformations are "learned" by memorizing sequences of a patch undergoing a transformation is a complete algorithm similar to the existing HMAX (in which S2 tunings are learned by sampling and memorizing random patches of images). A biologically more plausible learning rules would however be somewhat different: synapses would change as an effect of the inputs, in a sense trying to compress information. A very plausible learning rule is the associative Hebb's rule. The surprising obseravtion here is that the Hebb rule will force synapses at the level of the simple cells at each layer to shape the tuning of the receptive fields following the eigenvectors of the templatebooks. Thus this result shows that the receptive field at each layr should be determined by the transformations represented by the complex cells pooling at each layer.

In particular, I want to consider Oja's rule. It is not the only one with the properties we need but it is a simple rule and biologically plausible.

Oja's rule defines the change in presynaptic weights w given the output response y of a neuron to its inputs to be

$$\Delta \mathbf{w} \;=\; \mathbf{w}_{n+1} - \mathbf{w}_n \;=\; \eta\, y_n(\mathbf{x}_n - y_n \mathbf{w}_n) \tag{12}$$

where is the "learning rate". Notice that the equation follows (see Supp. Mat. [7]) from expanding to the first order Hebb's rule normalized to avoid divergence of the weights. Hebb's rule, which states in conceptual terms that "neurons that fire together, wire together", is written (In component form) as:

$$\Delta \mathbf{w} \;=\; \eta\, y(\mathbf{x}_n)\mathbf{x}_n \tag{13}$$

or

$$w_i(n+1) \;=\; w_i + \eta\, y(\mathbf{x}) x_i. \tag{14}$$

Hebb's rule has synaptic weights approaching infinity with a positive learning rate. In order to actually work the weights have to be normalized so that each weight's magnitude is restricted between 0, corresponding to no weight, and 1, corresponding to being the only input neuron with any weight. Mathematically, this requires a modified Hebb's rule:

$$w_i(n+1) \;=\; \frac{w_i + \eta\, y(\mathbf{x}) x_i}{\left(\sum_{j=1}^{m} [w_j + \eta\, y(\mathbf{x}) x_j]^p\right)^{1/p}} \tag{15}$$

of which Oja's rule is an approximation.

The version of Oja's rule that we consider is the version that applies to m neural units – in our case the D complex cells associated with the D rows of a templatebook:

$$\mathbb{W}_{k+1} = \mathbb{W}k + \mu_k [x_k - \mathbb{W}_k y_k] y_k^T \tag{16}$$

where $\mathbb{W}_{k+1} = [w_k(1) w_k(2) \cdots w_k(m)]$ is the weight matrix whose columns are the individual neuron weight vectors $w_k(i)$ and $y_k = \mathbb{W}_k^T x_k$ is the output vector of D elements.

The following *linking theorem* is a direct application to our setup of one of Oja's results:

Theorem 4. *If learning at the level of the receptive fields of the simple cells pooled by complex cells follow Oja's rule, then the receptive fields of the simple cells reflect the Principal Components of the associated templatebook.*

Thus this particular version of the architecture with this learning rule, links the spectral properties of $\mathbb{T}$ to the tuning of the simple units.

3.2. **Equivalent cells.** Figure 3 shows a cartoon of the key computational element of the model, corresponding to the *Neural Response* of the second Appendix.

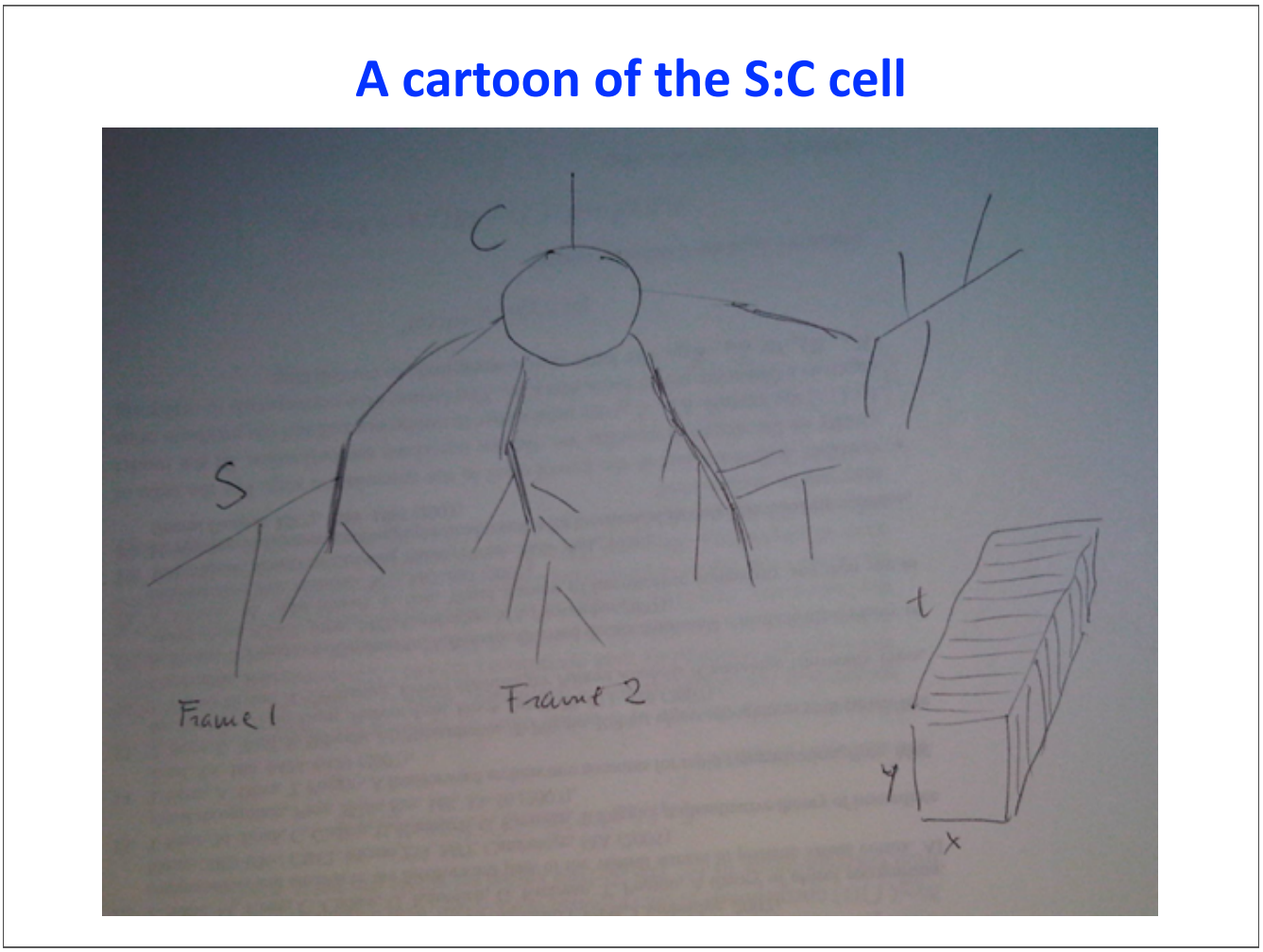

FIGURE 3. Cartoon of a SC cell with dendrites representing simple cells and the cell body performing complex-like pooling.

4. DISCUSSION

There are several key ideas in the theoretical framework of the paper. There are hypotheses and there are thorems.

(1) First, I conjecture that the sample complexity of object recognition is mostly due to geometric image transformations (different viewpoints) and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. The most surprising implication of the theory emerging from these specific assumptions is that the computational goals and detailed properties of cells in the ventral stream follow from *symmetry properties* of the visual world through a process of correlational learning. The obvious analogy is physics: for instance, the main equation of classical mechanics can be derived from general invariance principles. In fact one may argue that a Foldiak-type

rule determines by itself the hierarchical organization of the ventral stream, the transformations that are learned and the receptive fields in each visual area.
(2) Second, I assume that there is a hierarchical organization of areas of the ventral stream with increasingly larger receptive fields. I try to prove that increasing apertures determine a stratification of the invariances from translations to full affine in highest layers.
(3) The third idea is that transformations determine the spectral properties of samples of transformed images and thus of a set of templates recorded by a memory-based recognition architecture such as an (extended) HMAX.
(4) The fourth observation is that natural Hebbian learning algorithm ensure that spectral properties of the inputs determine corresponding receptive fields of simple cells.
(5) Finally, aggregation fucntions such as the max (in HMAX) ensures that signatures of images are invariant to affine transformations of the image and that this property is preserved from layer to layer.

The theory part of this paper start with this central computational problem in object recognition: identifying or categorizing an object after experience with a single image of it – or of an exemplar of its class. To paraphrase Stu Geman, the difficulty in understanding how biological organisms learn – in this case how they recognize – is not the usual $n \to \infty$ but $n \to 0$. The mathematical framework is inspired by known properties of neurons and visual cortex and deals with the problem of how to learn and discount invariances. Motivated by the Johnson-Lindenstrauss theorem, I introduce the notion of *signature* of an object as a set of similarity measurements with respect to a small set of template images. An *invariance lemma* shows that the stored transformations of the templates allow the retrieval of the signature of an object for any *linear transformation of it* such as an affine transformation in 2D. Since any transformation of an image can be approximated by local affine transformations (the *affine lemma*), corresponding to a set of local receptive fields, the invariance lemma provides a solution for the problem of recognizing an object after experience with a single image – under conditions that are idealized but likely to be a good approximation of reality. I then provide theorems characterizing how a hierarchical architecture may learn transformations in a sequence of stages and how the properties of the specific areas are determined by visual experience and continuous plasticity. At this point this section of the paper is more a research program than a theory, with more conjectures than theorems. It describes how the hierarchical architecture of the ventral stream with receptive fields of increasing size (roughly by a factor 2 from V1 to V2 and again from V2 to V4 and from V4 to IT) could implicitely learn during development different types of transformations starting with local translations in V1 to a mix of translations and scales and rotations in V2 and V4 up to more global transformations in PIT and AIT (the *stratification conjecture*). I characterize the spectral structure of various types of transformations that can be learned from images. The conjecture – to be verified with simulations and other empirical studies – is that in such an architecture the properties of the receptive fields in each area is mostly determined by the underlying transformations rather than the statistics of natural images.

The second part of the paper puts together the previous results into a detailed hypothesis of the plasticity, the circuits and the biophysical mechanisms that may subserve the computations in the ventral stream.

In summary, some of the broad predictions of this theory-in-fieri are:

- the type of transformation that are learned from visual experience depend on the size (measured in terms of wavelength) and thus on the visual area – assuming that the aperture size increases with area;
- the mix of transformations learned determine the properties of the receptive fields – oriented bars in V1+V2, radial and spiral patterns in V4 up to class specific tuning in AIT (eg face tuned cells);
- pure examples of isolated transformations (translations, rotations, expansions) correspond to the perception of Glass patterns and correspond to a simple model of learning transformations;
- class-specific modules – such as faces, places and possibly body areas – should exist in IT to process images of object classes;
- the output of the ventral stream is a *signature* to be used as key for an associative memory (or of a vector-valued classifier): properties of the code and bound on the storage and retrieval capabilities of the memory can be characterized.
- during evolution, areas above V1 should appear at later times, reflecting increasing object categorization abilities.

Some remarks are:

- The invariance of a layer of complex cells is a metaphor for the symmetry represented in that layer; the tuning of the corresponding simple cells are a metaphor for the tuning of the layer
- the tuning follow from the invariance of the complex cells
- the invariance of the complex cells follows from the size of the receptive field of the complex cells in that layer
- From the assumption of a hierarchy of aereas with receptive fields of increasing size the theory predicts that the size of the receptive fields determines which transformations are learned during development and then factored out during normal processing; that the transformation represented in an aerea determines the tuning of the neurons in the aerea; and that class-specific transformations are learned and represented at the top of the hierarchy.

5. APPENDICES

5.1. **Appendix I: empirical evidence from the horses-dogs challenge.** The conjecture of section 1.1 implies that recognition should be easy if image transformations are factored out. The dog-house challenge of Figure 4 tests this prediction. The task is to categorize correctly dogs vs horses with a very small number of training examples (eg small sample complexity).

All the 300 dogs and horses are images obtained by setting roughly the viewing parameters – distance, pose, position. With these normalized images,there is no significant difference between running the classifier directly on the pixel representation versus using a more powerful set of features (the C1 layer of the HMAX model).

insert more recent figures from Joel

5.2. **Appendix II: mathematics of the invariant neural responses.** We provide a mathematical description of the neural response architecture [14] of Figure 4, which is designed to be robust to transformations encoded implicitly in sets of templates. Robustness is achieved by mean of suitable pooling operations across the responses to such templates. The setting we describe is a modification of the one introduced in [14].

5.2.1. *Framework.* We start giving the basic concepts and notations describing the framework we consider. *Architecture Elements* The new neural response is defined by an architecture composed of the following elements.

- A finite number of nested sets $p_1 \subset p_2 \subset \cdots \subset p_n$, that we call patches.
- A family of function spaces defined on each patch

$$(\mathrm{Im}(p_i))_{i=1}^n, \quad \text{where} \quad \mathrm{Im}(p_i) = \{x \mid x : p_i \to [0,1]\}, \quad i = 1, \ldots, n.$$

- A family of finite sets of maps from a patch to the next larger one,

$$(H_i)_{i=1}^{n-1}, \quad \text{where} \quad H_i = \{h \mid h : p_i \to p_{i+1}\},$$

that we call decomposition maps. The name is justified by the observation that H_i describe how a function $x \in \mathrm{Im}(p_{i+1})$ can be decomposed in a set of functions $x \circ h$, $h \in H_i$, with smaller domain, namely a set of *parts*.

5.2.2. *Tuning Function.* A tuning function $K : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ is given, which is a reproducing kernel Hilbert space. The tuning function can be naturally restricted to $\mathbb{R}^b \times \mathbb{R}^b$, with $b \leq d$. The two main examples of tuning function we have in mind are the Gaussian $K(x,x') = \exp -\gamma \|x - x'\|^2$ and the normalized inner product $K(x,x') = \frac{\langle x,x' \rangle}{\|x\|\|x'\|}$, where $\langle \cdot, \cdot \rangle$, $\|\cdot\|$. are the inner product and norm in $\mathbb{R}^d$.

5.2.3. *Families of Invariance Sets.* A last crucial ingredient is needed to define the generalized neural response. A family of sets whose elements are themselves sets of functions, that is

$$(V_i)_{2=1}^n \quad \text{where} \quad V_i = \{v \mid v = \{t \mid t \in \mathrm{Im}(p_i)\}\}, \quad i = 2, \ldots, n.$$

We assume that $|V_i| \leq d$, and $|v| \leq d$ for $v \in V_i$ and all $i = 2, \ldots, n$. Each element v of a set V is called an invariance set.

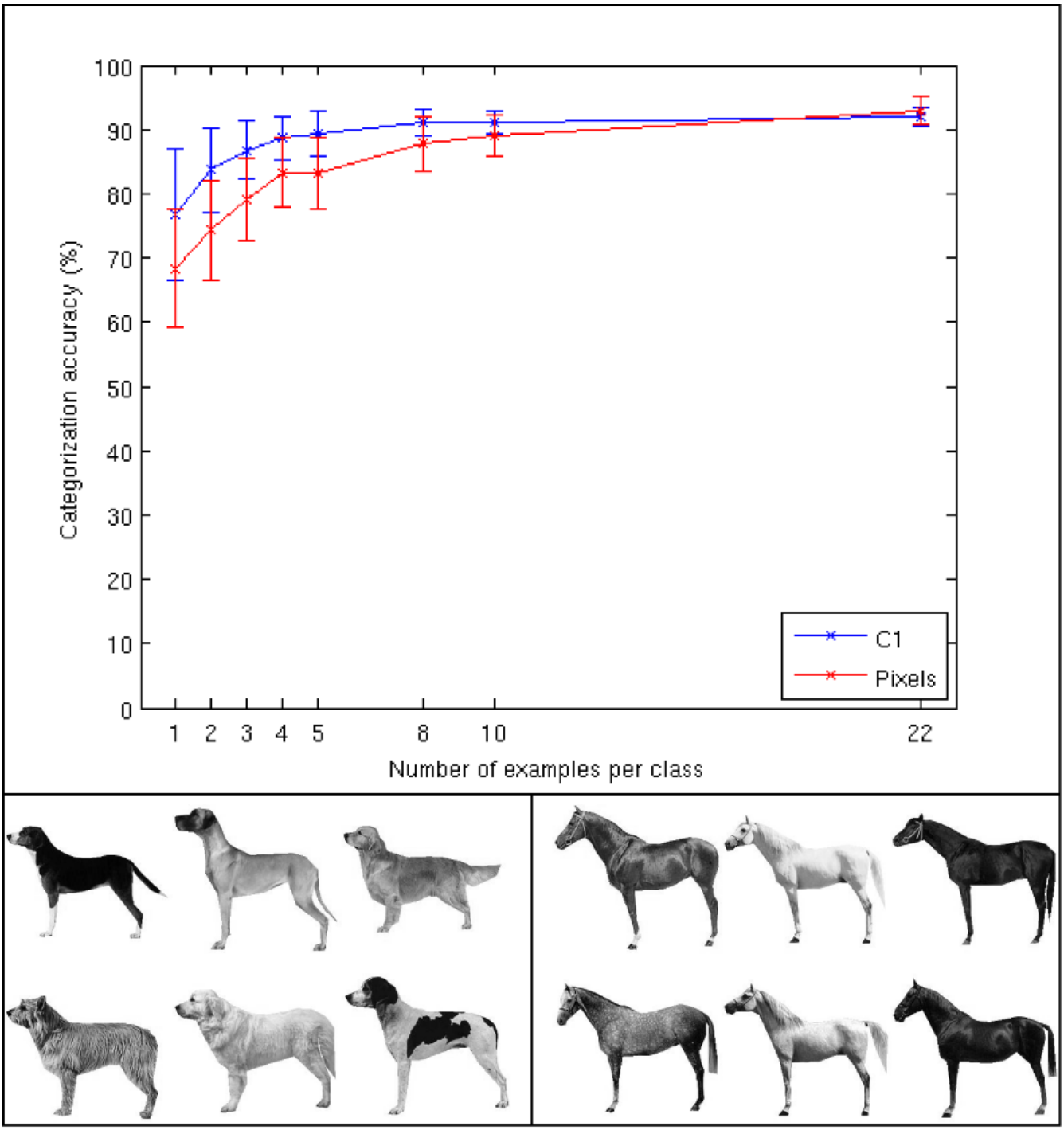

FIGURE 4. Images of dogs and horses, 'normalized' with respect to image transformations. A regularized least squares classifier (linear kernel) tested on more than 150 dogs and 150 horses does well with little training. Error bars represent +/- 1 standard deviation computed over 100 train/test splits. This presegmented image dataset was provided by Krista Ehinger and Aude Oliva.

5.2.4. *New Neural Response Definition.* The definition of the generalized invariant neural responses is the following.

Definition 2. *Given an initial neural response $N_1 : \mathrm{Im}(p_1) \to \mathbb{R}^p$, $p \leq d$, the m-layer neural response $N_m : \mathrm{Im}(p_m) \to \mathbb{R}^{|V_m|}$, for $m = 2, \ldots, n$, is defined as*

$$N_m(x)(v) = \max_{t \in v} \left\{ \sum_{h \in H} K(N_{m-1}(x \circ h), N_{m-1}(t \circ h)) \right\} \tag{17}$$

with $x \in \mathrm{Im}(p_m)$, $h \in H_{m-1}$, $v \in V_m$.

5.2.5. *Learning.* The interpretation of the above model that suggests how the invariance can be learned from data. Having in mind problems in computational vision, we think of $\mathrm{Im}(p_i)_{i \geq 1}$ as images of increasing size.

The patches can be thought of as squared domains centered around the origin. The decomposition maps describe how an image can be decomposed into (possibly overlapping) image patches. In the above construction an invariance set $v \in V_m$ is often an ordered set of images. In practice this set can be obtained from a video sequence v so that, if $v = \{t_1, \ldots, t_p\}$ then $t_1, \ldots, t_p$ correspond to frames at successive instants of time. The recording of sets of video sequences is the learning phase of the above model.

REFERENCES

[1] G. Hinton and R. Memisevic. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 22:14731492, 2010.

[2] J. Leibo, J. Mutch, and T. Poggio. How can cells in the anterior medial face patch be viewpoint invariant? *MIT-CSAIL-TR-2010-057, CBCL-293; Presented at COSYNE 2011, Salt Lake City*, 2011.

[3] J. Leibo, J. Mutch, and T. Poggio. Learning to discount transformations as the computational goal of visual cortex? *Presented at FGVC/CVPR 2011, Colorado Springs, CO.*, 2011.

[4] J. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio. Invariant Recognition of Objects by Vision. *CBCL-291*, 2010.

[5] J. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio. Learning Generic Invariances in Object Recognition: Translation and Scale. *MIT-CSAIL-TR-2010-061, CBCL-294*, 2010.

[6] J. Leibo, J. Mutch, S. Ullman, and T. Poggio. From primal templates to invariant recognition. *MIT-CSAIL-TR-2010-057, CBCL-293*, 2010.

[7] T. Poggio. The computational magic of the ventral stream: Supplementary Material. *CBCL Internal Memo*, 2011.

[8] R. P. Rao and R. D. L. Learning lie groups for invariant visual perception. 1999.

[9] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, Nov. 1999.

[10] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3(11), 2000.

[11] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *CBCL Paper #259/AI Memo #2005-036*, 2005.

[12] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–6429, 2007.

[13] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–426, 2007.

[14] S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. Mathematics of the neural response. *Foundations of Computational Mathematics*, 10(1):67–91, 2010.

[15] C. F. Stevens. Preserving properties of object shape by computations in primary visual cortex. *PNAS*, 101(11):15524–15529, 2004.

[16] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 992–1006, 1991.